Math 574: Bayesian Computational Statistics
Predicting Hall of Fame Coaches
in Professional Sports

Aleksei Sorokin — asorokin@hawk.iit.edu — A20394300

May 1, 2020

1 Introduction

This paper builds logistic regression models in a Bayesian framework to predict if a professional sports coach will be inducted into the hall of fame. Specifically, we pull coaching data from the National Football League, National Basketball League, Major League Baseball, and National Hockey League to build several logistic regression models.

1.1 Data Loading and Cleaning

Football, basketball, baseball, and hockey data for this project were sourced from SportsReference.com [1]. The common attributes between the four sports datasets are Name, Games Coached, Win-Loss Percentage, Playoff Win Loss Percentage, Conference Championships, Championships, and Hall of Fame Status. Individual datasets contained unique sport specific attributes that were not used in this analysis. For example, the baseball coaches dataset contained an attribute of how many all star games the coach had been selected to manage. While such information would create a more robust, and probably a more accurate, model, this work focuses on predicting hall of fame status for coaches across all sports. With this in mind, it became important to regularize certain common attributes provided from each dataset.

The number of games a coach lead was divided by the number of games in a regular season of the sport. This is necessary as different sports have different number of games in a season i.e. baseball has 162 games in a season while football has only 16. The total years the coach lead a team was dropped in favor of games played divided by season length. The later statistic accounts for coaches who did not complete the entire season as the head coach. In a similar way to regularizing total number of games coached, the number of playoff games coached was divided by the minimum number of playoff games that must be won to win the championship in that sport. For example, a basketball team that has made the playoffs must win 4 best-of-7 series in order to win the championship. Therefore, the number of playoff games a basketball coach lead was divided by 16 to regularize the score.

After filling not-available data with 0 values, regularizing the data, and selecting explanatory variables, the combined dataset of all coaches contained attributes described by the following data dictionary.

- **N**: Name
- **GR**: Games relative to season length
- **WP**: Win-loss percentage
- **PGR**: Playoff games relative to minimum games needed to win championship
- **PWP**: Playoff win-loss percentage
- **CC**: Conference championships (won qualifier to play in championship)
- **C**: Championships
- **S**: Sport i.e. football, basketball, baseball, or hockey
- **HOF**: Hall of fame status. 1 - made HOF, 0 - not

1.2 Data Analysis

Data analysis was preformed on the correlation between predictor variables. As shown by Figure 1, there exist strong positive correlations between games (relative) and playoff games (relative), playoff win percentage, conference championships, and championships. These correlations make sense as the more seasons a coach leads a team, the more likely he is to play in the playoffs and win conference and national championships. More surprising is the strong positive relationship between games and playoff win loss percentage. This relationship suggests that the longer a coach heads a team, the more likely he is to win in the playoffs. Perhaps this enforces how important playoff experience is not only for the players, but for the coaches as well. Figure 1 also shows a strong positive correlation between playoff games and playoff win-loss percentage, conference championships, and national championships. Again, this correlation makes sense as coaches who participate in more playoff games have experience which helps to bolster their chance of winning and raise the odds of them winning championships for their team/organization.

Data analysis was performed on the underlying distribution of common attributes. For each numeric predictor variable itemized in the previous subsection, a histogram-boxplot was created and top coaches in that category were aggregated. Figure 2 shows and example of the plot and aggregated top-10 dataframe created for games relative (roughly number of years coached). Such analysis leads to distributions with large numbers of outliers. However, this is expected as coaches who do not preform well often do not maintain their head coach position for many games, thus making coaches with longer careers appear as outliers in the dataset. Similarly, most coaches never win a championships, so those who win even one or two championships are considered outliers in the datasets. Another common form of outlier are coaches who had success in such a small number of games their stats reflect a much better performance than achieved in their actual coaching career. In a concrete example, there are over 10 coaches who have a perfect 1.00 win-loss percentage. However, none of these coaches have lead a team for more than a fifth of a season. Therefore, their statistics are not stable and do not accurately reflect their success as a professional coach. One idea to remedy these "one-game-wonders" was to drop coaches who had coached for less than a year's worth of games. The problem with this approach is that there are many coaches who have been nominated to the hall of fame despite being a head coach for less than a year. Thus our data cannot reflect the entire career of a professional coach, as further explained in the following subsection.

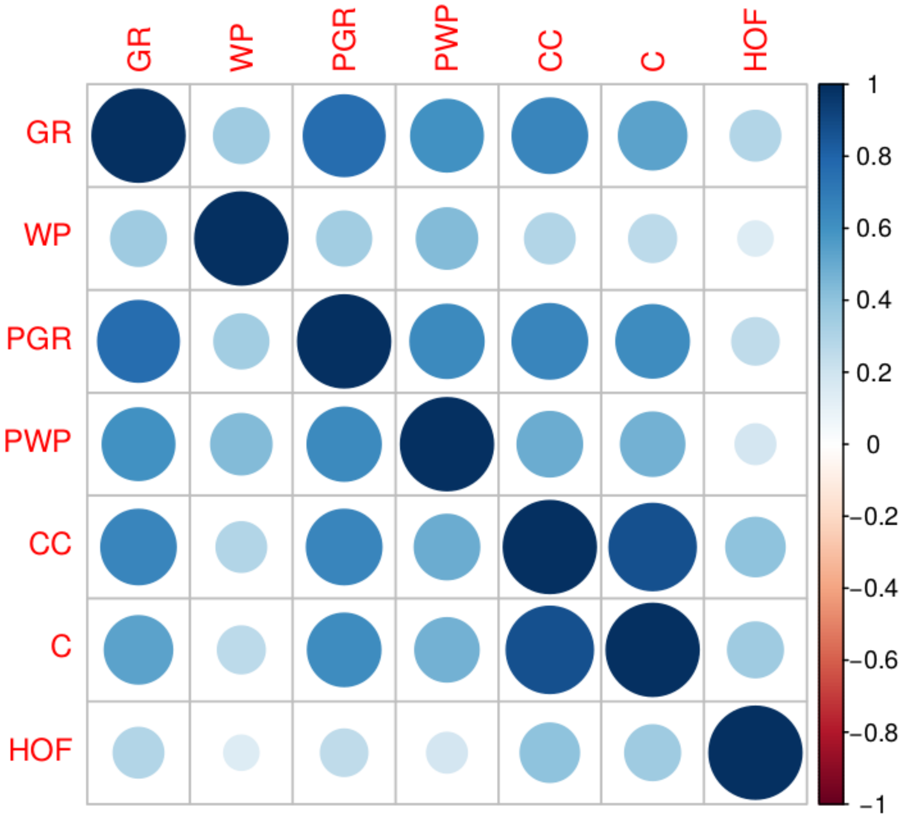
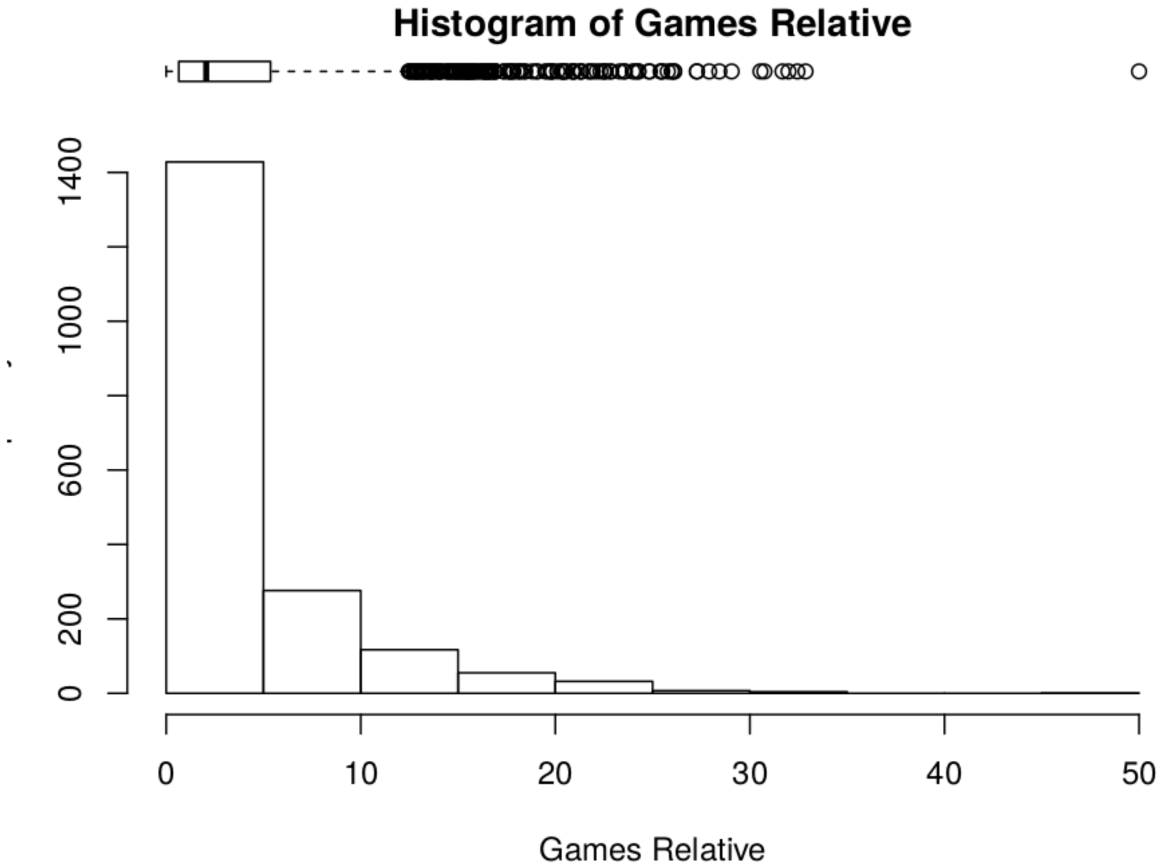


Figure 1: Covariance between common coaching attributes used as explanatory variables in modeling.

1.3 Problems with Data

While statistical data can summarize the numerical success of a coaches career, a coaches nomination to the hall of fame is influenced by many other factors. One such factor, as hinted at above, is a coaches playing career. Barney Stanley was an excellent Hockey player who played from 1914 to 1929 yet only coached one



```
## Top 10 coaches by Games Relative
##                  N       GR    WP       PGR   PWP CC C HOF          S
## 403    Connie Mack 47.87037 0.486  3.909091 0.558  9 5   1   baseball
## 1856 Tony La Russa 31.46296 0.536 11.636364 0.547  6 3   1   baseball
## 754   George Halas 31.06250 0.682  3.000000 0.667  6 6   1   football
## 562       Don Shula 30.62500 0.677 12.000000 0.528  6 2   1   football
## 1277 Lenny Wilkens 30.32927 0.536 11.125000 0.449  2 1   1 basketball
## 1145   John McGraw 29.43827 0.586  4.909091 0.481 10 3   1   baseball
## 560     Don Nelson 29.24390 0.557 10.375000 0.452  0 0   1 basketball
## 236      Bobby Cox 27.82716 0.556 12.363636 0.493  5 1   1   baseball
## 287   Bucky Harris 27.22222 0.493  1.909091 0.524  3 2   1   baseball
## 1102     Joe Torre 26.72222 0.538 12.909091 0.592  6 4   1   baseball
```

Figure 2: Histogram-boxplot for games relative to season length. Note that Connie Mack coached for 50 years although he coached the game-equivalent of less than 48 complete seasons.

year of professional hockey in the 1927-1928 season. His coaching record cannot account for his stats as a player as no reasonable model should predict a coach with less than half a year of coaching experience, who had less than a 25% win-loss percentage, and who coached 0 playoff games to be nominated to the hall of fame. Yet his professional career as a player warrants the nomination. Another problem with data that was touched on above are coaches with short careers who accumulated an unstable statistical record. One example is Chuck Drulis who won the only two games he coached in the NFL, thus giving him with an outstanding 100% win-loss percentage. Finally, it is important to acknowledge that statistics cannot tell the entire story of a professional coach. A coach's impact on the players, league, and community is also taken into account when making the decision of weather or not a coach is worthy of induction into the hall of fame.

All of this is to say that the models developed in the following section are only as good as the underlying data, which in this case is susceptible to unlikely outliers and anomalies resulting from the inconsistent and unpredictable nature of professional coaching.

2 Simple Bayesian Logistic Regression

This section gives a simple example of a Bayesian logistic regression model. Rather than using all common predictors as is explored in the following section, this section builds a model with only GR (games relative) predicting a coaches hall of fame status. The two parameters of interest are the intercept, β_0, and the coefficient of the GR predictor, β_1. By using the built in glm function in R with a binomial family, the optimal coefficients were found to be -2.448 and .01118 for β_0 and β_1 respectively. In a Bayesian context, the non-informative prior $p(\beta_0, \beta_1) \propto 1$ leads to similar coefficient estimates as the logistic regression fit with maximum likelihood estimation. However, the Bayesian context allows for greater insight about the posterior distributions and allows for simulations from the posterior distribution using marginal densities and the inverse CDF method. Figure 3 shows a contour plot based on posterior simulations of β_0 and β_1. Using these posterior simulations, a 90% confidence interval for β_0 was found to be (-2.63,-2.27) and for β_1 was found to be (1.10,1.14). These intervals capture the MLE estimates of β_0 and β_1, thus confirming the similarity of the two methods.

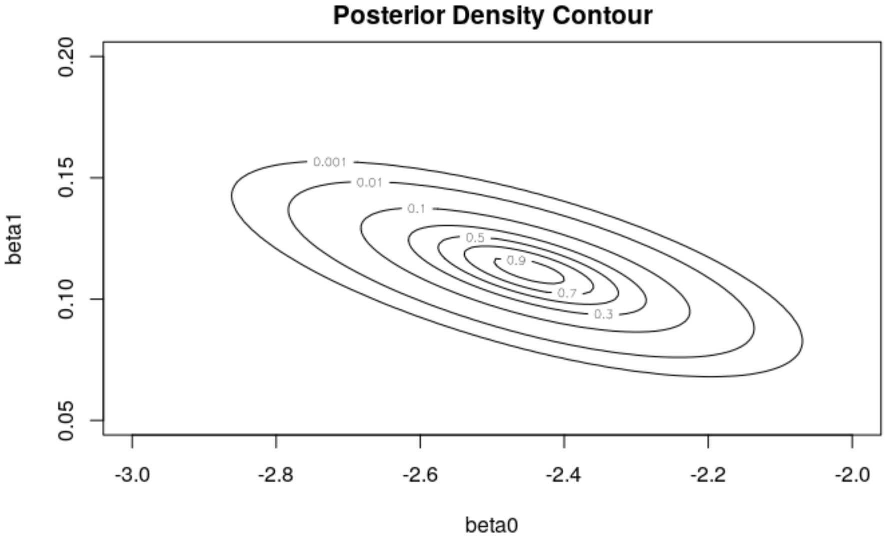


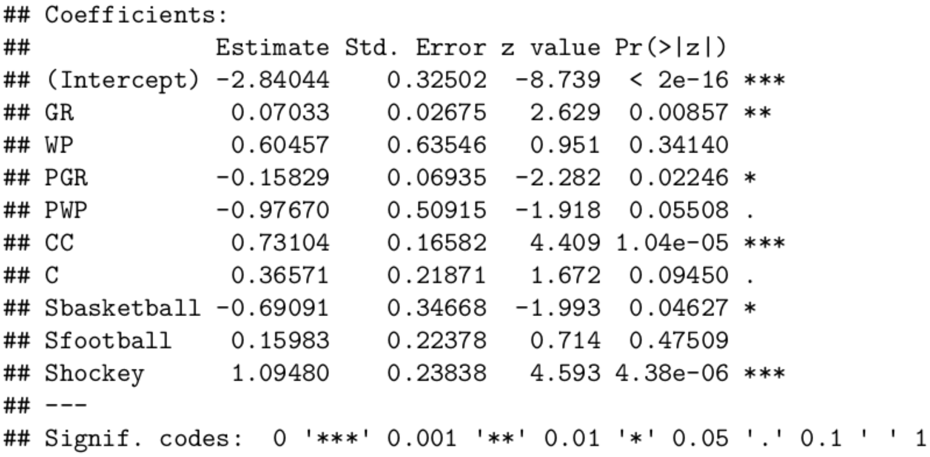
Figure 3: Contour plot of posterior distribution for β_0 and β_1.

3 Models

Three separate models were fit to the dataset. The first model was a standard logistic regression model with sport treated as a categorical variable. The second model was a logistic mixed-effect model fit with maximum-likelihood estimation where sport was the random effect. The third model was a logistic mixed-effect model in the Bayesian framework where sport was again used as the random effect parameter. Each of the models was fit on a training dataset and tested on a disjoint testing dataset. 80% of the dataset of coaches from all sports was used for training and 20% was used for testing. To maintain a consistent imbalance of hall of fame coaches across the training and testing datasets, 80% of hall of fame coaches were put in the training dataset and the remaining 20% were used in the testing dataset. The resulting training dataset consisted of 204 hall of fame coaches and 1331 non-hall of fame coaches while the testing dataset consisted of 52 hall of fame coaches and 333 non-hall of fame coaches. It is important to note the heavy imbalance in the dataset of hall of fame coaches to non-hall of fame coaches. This imbalance caused the models to overestimate the number of non-hall of fame coaches, thus resulting in low recall of all three models. The parameters and results from each model are discussed in more depth in the following subsections

3.1 Logistic Model without Random Effect

The most basic logistic regression model was fit with fixed effect on GR (Games Relative), WP (Win Percentage), PGR (Playoff Games Relative), PWP (Playoff Win Percentage), CC (Conference Championships), C (Championships), and S (Sport). For this model, sport was treated as a categorical (factor) variable with four categories: football, basketball, baseball, and hockey. The resulting model fit with coefficients in Figure 4 indicated that the intercept, GR, PGR, CC, and S were the coefficients with less than a .1 p-value for the null hypothesis the coefficient is 0. In other words, The number of games, both regular season and playoff, the number of championship, and the sport the coach was involved in are all significant in determining weather or not a coach should be in the hall of fame. The coefficient for PGR is negative, indicating that the more playoff games a coach is involved in, the less likely they are to be in the hall of fame. This does not hold true to reality where we expect coaches with more playoff experience to generally be considered more successful than coaches who have made less playoff games. The significantly positive coefficient on hockey shows that is may be easier to get into the hockey hall of fame than say the football, baseball, or basketball hall of fame. In fact, the basketball sport indicator has a significantly negative coefficient, indicating it may be even more difficult to make the basketball hall of fame than any of the other three sports.



```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -2.84044    0.32502  -8.739  < 2e-16 ***
## GR           0.07033    0.02675   2.629  0.00857 ** 
## WP           0.60457    0.63546   0.951  0.34140    
## PGR         -0.15829    0.06935  -2.282  0.02246 *  
## PWP         -0.97670    0.50915  -1.918  0.05508 .  
## CC           0.73104    0.16582   4.409 1.04e-05 ***
## C            0.36571    0.21871   1.672  0.09450 .  
## Sbasketball -0.69091    0.34668  -1.993  0.04627 *  
## Sfootball    0.15983    0.22378   0.714  0.47509    
## Shockey      1.09480    0.23838   4.593 4.38e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4: Resulting fixed effect parameters from fitting training data to a logistic regression model with sport as a categorical variable.

3.2 Logistic Model with Sport Random Effect

The second model was was fit with fixed effects GR, WP, PGR, PWP, CC, and C and had mixed effect for S (Sport) using the *lmer* function from the *lme4* package [2]. Sport was selected as the mixed effect parameter as there may be significant differences in criterion for making the hall of fame depending on the sport. For example, since baseball is the oldest of the four sports and coaches generally have a higher GR, one may need a better resume to make the baseball hall of fame than say the football or hockey hall of fame. Compared to the logistic regression model without random effect, the coefficients of this model (see Figure 5) are closer to 0 and have much smaller standard error. For instance, the intercept of the first model is around -2.8 while the intercept of this model is around .066. While the coefficients are closer to 0 in this model, PGR and PWP are still negative, indicating that playoff experience hurts a coaches chances of making the hall of fame. This is an interesting trend as it goes directly against the achievement of a coach making the playoffs. One possible explanation is that the majority of coaches have 0 PGR and 0 PWP, meaning they have been in 0 playoff games in their career. This skews the playoff data and perhaps contributes to the negative coefficients. This model fits the Sport random effect with a intercept of around .004 and with a residual of .095. While the intercept and residual are relatively small, the resulting increased flexibility of this model makes for fixed effect parameters with smaller coefficients. Figure 5 shows the resulting coefficients of this model.

```
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.5487 -0.4286 -0.2610 -0.1408  3.2702
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  S        (Intercept) 0.00429  0.0655
##  Residual             0.09513  0.3084
## Number of obs: 1535, groups:  S, 4
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)  0.066260   0.040566   1.633
## GR           0.008219   0.002676   3.072
## WP           0.054907   0.056677   0.969
## PGR         -0.018178   0.006785  -2.679
## PWP         -0.088286   0.048663  -1.814
## CC           0.103399   0.017237   5.999
## C            0.031050   0.021654   1.434
```

Figure 5: Resulting mixed and fixed effect parameters from fitting training data to a logistic regression model with sport as a random effect variable.

3.3 Logistic Model in Bayesian Framework with Sport Random Effect

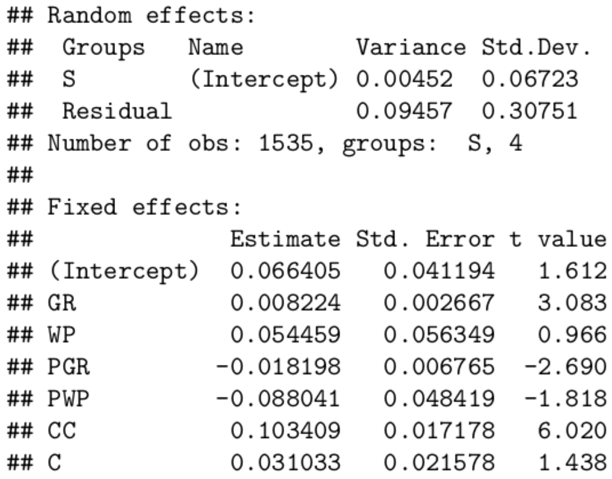
The final model is a Bayesian logistic regression model with sport used as group effect variable. The fixed effects were given a normal prior, the mixed effect (sport) was given a inverse Wishart prior, and the residual was given a Gamma prior. To execute the model the *blme* function was used from the *blmer* package [3]. This package greatly simplifies the Bayesian logistic mixed-effect model fitting process and gives a nice summary of parameters, coefficients, and details of the fit. It should be noted that many other priors were tried for the fixed, mixed, and residual parameters, but all gave almost identical coefficients and estimates. The resulting coefficients for both the fixed and mixed effects (found in Figure 6) are almost identical to the previous model with the same structure fitted with maximum likelihood estimation rather than starting from Bayesian priors as we do in this framework. The similarity in coefficients are due to the large amount of data and the non-informative priors. With a total of over 1,500 coaching records in the training dataset, the data dominates the non-informative priors and results in estimates derive in the same way as the non-Bayesian logistic mixed model. Figure 6 shows the resulting coefficients of this model.

3.4 Comparison of Model Prediction Metrics

Due to the binary hall of fame label we are trying to predict, the appropriate metrics that were recorded are accuracy, precision, and recall. To collect these metrics the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values were recorded based on predictions from each model.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$



```
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  S        (Intercept) 0.00452  0.06723
##  Residual             0.09457  0.30751
## Number of obs: 1535, groups:  S, 4
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)  0.066405   0.041194   1.612
## GR           0.008224   0.002667   3.083
## WP           0.054459   0.056349   0.966
## PGR         -0.018198   0.006765  -2.690
## PWP         -0.088041   0.048419  -1.818
## CC           0.103409   0.017178   6.020
## C            0.031033   0.021578   1.438
```

Figure 6: Resulting mixed and fixed effect parameters from fitting training data to a logistic regression model with sport as a random effect variable in a Bayesian framework.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Due to the almost identical coefficients of both the non-Bayesian and Bayesian Logistic mixed models, the two models made the same predictions on all coaches in the test dataset. Therefore, we only need to compare metrics between the standard logistic model and mixed effect logistic model. The accuracy of the logistic and mixed effect logistic models are both at 88%. The precision of the standard logistic model is around 71% compared to around 90% for the mixed effect models. However, the standard logistic model has around 23% recall compared to only 17% for the mixed logistic models. All three models are fairly accurate, but the majority of this accuracy can be contributed to correctly predicting coaches to not make the hall of fame. This relationship is evident in the low recall metrics which indicates our models have many false negatives: the models are not correctly identifying coaches that should be in the hall of fame. However, when our models do predict a coach to make the hall of fame, they are fairly accurate, as indicated by the high precision values.

These results are unimpressive yet unsurprising. As discussed in the Data Analysis section, a coaches induction to the hall of fame is based on many factors not captured by the data available. Taking a closer look at the models false negatives, we see evidence of the problems discussed before the models were fit. Take Alan Trammell for instance. One would not expect a model to predict a coach with around 3 years of experience, less than a 40% win loss percentage, and who has never coached in a playoff game to be in the hall of fame. Yet he was inducted based on his 20 years of playing shortstop for the Detroit Tigers before of his brief coaching career. Although there are some coaches whose greatness is not captured in the data, there are other coaches who preform well in all statistical categories yet still fail to be selected for the hall of fame by the models. Take at Bill Cowher for instance. Bill was elected to the hall of fame based on his 15 years of head coach experience, over 20 playoff games, and 2 conference championships. Yet all three models gave him less than a 30% chance of making the hall of fame. Overall, the imbalanced data and lack of external explanatory variables resulted in models that fail to consistently identify coaches who should be in the hall of fame.

3.5 Conclusion

Overall, there were minimal differences between standard logistic regression using sport as a categorical variable versus using sport as a mixed effect variable. There were even smaller differences between fitting an imperial mixed effect model and the mixed effect model in the Bayesian framework. While standard logistic regression yielded a higher recall, the mixed models produced higher precision and had coefficients with smaller standard errors. The low recall metrics for all three models indicates a failure to create models that accurately predict if a coach will make the hall of fame. However, such shortcomings are better understood when considering the many factors that contribute to a coaches hall of fame nomination that could not be captured in the data.

4 Future Work

One major limitation was the use of the *blme* function from the *blmer* package [3]. This package has sparse documentation and provides surprisingly little information on obtaining posterior distributions despite the framework of the models it builds. While many different prior distributions were tried, the data always seemed to dominate any combination of priors to result in the same coefficients and predictions. For this project I opted to include more predictor variables and a mixed effect variable rather than using say one predictor variable and one mixed effect predictor. This choice of including many fixed effect variables necessitated the use of a higher level package. In the future, it may be interesting to include only one fixed effect, say Games Relative (GR), and only the Sport mixed effect. Using a smaller subset of variables would allow one to write simulations manually and thus do much more analysis on the posterior distributions than the *blmer* package [3] currently allows. That being said, my experimentation with combinations of different fixed effect variables, priors, and mixed effects did not yield more encouraging metrics or more insightful posteriors.

5 Code Organization

The code is organized into R-Markdown files in a GitHub repository [4]. PDF renderings of these markdown files are maintained in *root/code_pdf/*.

1. *root/load_clean_data.Rmd*: Load data from sports-reference.com [1] for coaches in football, basketball, baseball, and hockey. Performs cleaning of each dataset, including renaming columns, filling NA values with 0, and binding common columns from all four datasets.

2. *root/explore_data.Rmd*: Extracts insights about the data, such as the total number of hall of fame (256) and non-hall of fame (1920) coaches. This notebook also plots the correlation between predictor variables and makes histogram-boxplots of each variable. At the end of the notebook is a brief list of potential problems with the data identified before modeling was preformed.

3. *root/bayesian_simulation.Rmd*: Fits a logistic regression model with only GR (Games Relative) as a predictor. Performs β_0 and β_1 simulation based on non-informative prior.

4. *root/modeling.Rmd*: Split data into training and testing datasets. Creates standard logistic regression model, logistic mixed model, and Bayesian logistic mixed model. Before the Bayesian logisitc mixed model is fit, each predictor variable was fit to a distribution using the *fitdist* function from the *fitdistrplus* package [5]. See Figure 7 for an example. Finally, metrics were aggregated and recorded for each of the three models.

6 Computational Extras

Throughout the semester I have continued to work on QMCPy [6], a Quasi-Monte Carlo framework in Python, with Fred Hickernell, Sou-Cheng Terrya Choi, Michael McCourt, and Jagadeeswaran Rathinavel. This research will continue over the summer with the continued support of SigOpt and the College of Science Summer Stipend.

After working on predicting Hall of Fame coaches with various statistical models, I began to work on implementing various sampling techniques discussed in class into the QMCPy framework. Specifically, QMCPy now includes inverse CDF sampling, acceptance-rejection sampling, and importance sampling. A Jupyter Notebook on sampling from various measures (sampling_measures.ipynb) demos implementations and use-cases of the above three sampling techniques. Specifically, this notebook shows examples of inverse

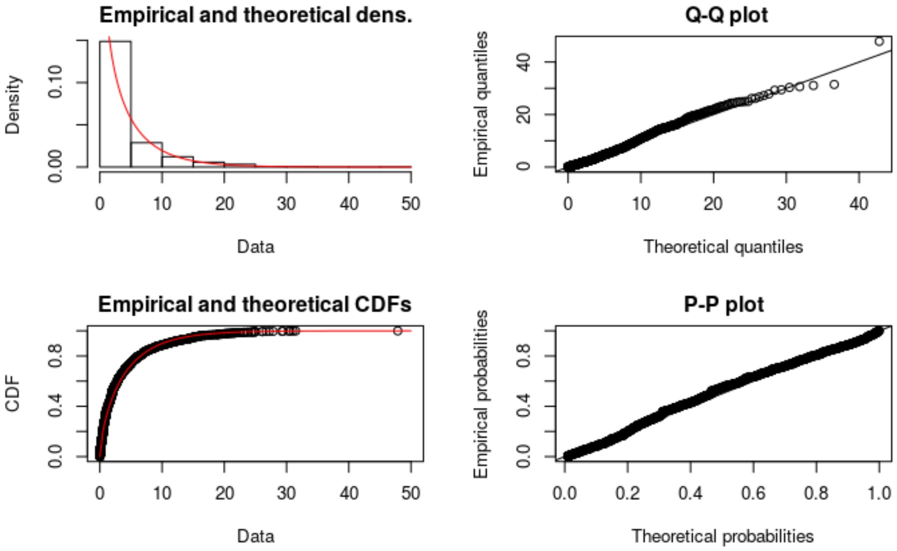


Figure 7: Fitting a distribution to the Games Relative (GR) predictor.

CDF sampling from the exponential and cauchy distributions (see Figure 8), acceptance rejection sampling from a trapazoidal pdf and a Bayesian example, and importance sampling from a uniform pdf over the quarter unit circle. Another Jupyter Notebook is specifically dedicated to importance sampling (importance_sampling.ipynb) and demos a game example and Brownian Motion drift (see Figure 9). Renderings of these notebooks and thorough documentation is maintained on the QMCPy homepage in ReadTheDocs.io. Over the summer I plan to implement Markov Chain Monte Carlo sampling techniques such as the Gibbs sampler, Metropolis algorithm, and Metropolis-Hasting algorithm. In addition, we are working to implement Jagadeeswaran's PhD work on Bayes Cubature algorithms for the lattice and Sobol low discrepancy sequences.

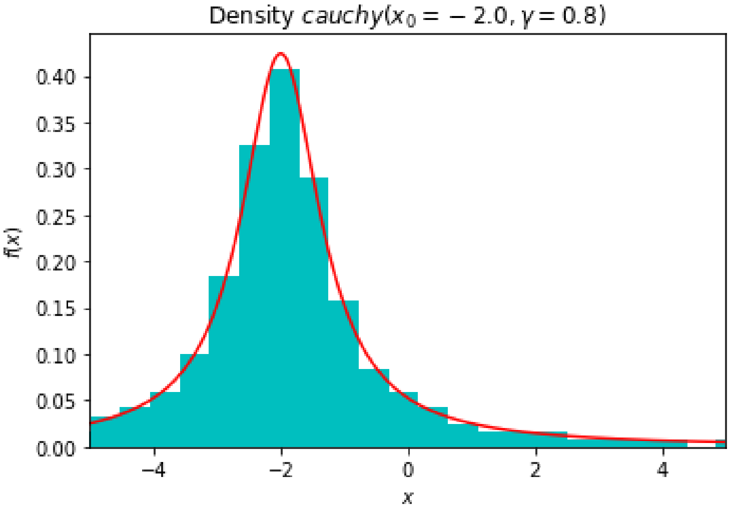


Figure 8: Sampling from Cauchy distribution using inverse CDF method.

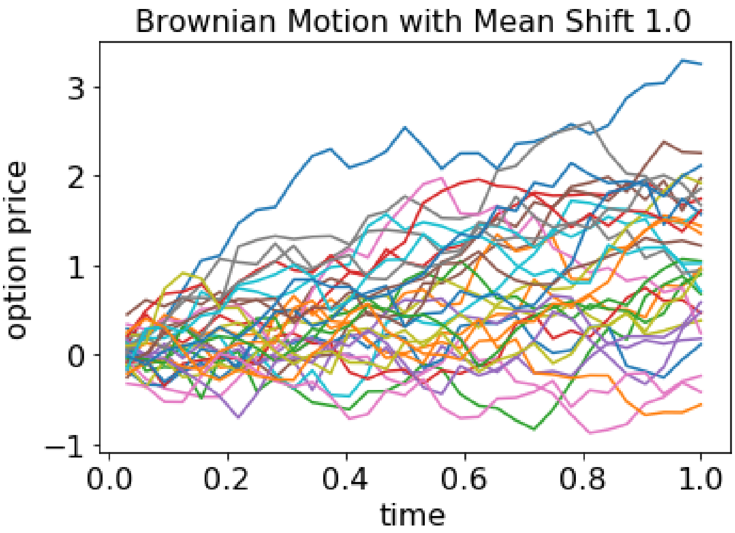


Figure 9: Brownian Motion with importance sampling implemented with the mean shift parameter.

References

[1] “Sports reference.” https://www.sports-reference.com/.

[2] D. Bates, M. Mächler, B. Bolker, and S. Walker, “Fitting linear mixed-effects models using lme4,” *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[3] Y. Chung, S. Rabe-Hesketh, V. Dorie, A. Gelman, and J. Liu, “A nondegenerate penalized likelihood estimator for variance parameters in multilevel models,” *Psychometrika*, vol. 78, no. 4, pp. 685–709, 2013.

[4] “Math574fp_coachinghof github repository.” https://github.com/alegresor/Math574FP_CoachingHOF.

[5] M. L. Delignette-Muller and C. Dutang, “fitdistrplus: An R package for fitting distributions,” *Journal of Statistical Software*, vol. 64, no. 4, pp. 1–34, 2015.

[6] “Qmcsoftware github repository.” https://github.com/QMCSoftware/QMCSoftware.